Commentaries on *Drug & Alcohol Findings* Research Analysis of Article on PROSPER Effects at 6½ Years Past Baseline

First Commentary on *Research Analysis of Article on PROSPER Effects on Substance Misuse through 6½ years Past Baseline* (Submitted to *Drug & Alcohol Findings* on 10-8-17)

(Richard Spoth[a], Cleve Redmond[a], Chungyeol Shin[a], Mark Greenberg[b], Mark Feinberg[b], & Lisa Schainker[a])

[a]Partnerships in Prevention Science Institute, Iowa State University, 2526 N Loop Drive, Suite 2400, Ames, IA 50010, USA
[b]Prevention Research Center, The Pennsylvania State University, Henderson Blvd S, Room 109, University Park, PA 16802, USA

We appreciate the opportunity to respond to key points made in the *Findings* commentary on the article referenced in our title.

The *Findings* commentary addressed two types of issues, concerning both the validity of the findings (whether the interventions were "really shown to be effective") and the replication of findings (whether the system would "work elsewhere"). We respond first and primarily to the validity issues raised, proceed to some comments on the replicability question, and close with some conclusions concerning some related, broader issues.

In general, we commend *Findings* for raising and discussing important issues relevant to the interpretation of reported findings from the PROSPER and other prevention trials. That said, we encourage a more balanced and comprehensive consideration of the findings from the reviewed article, in the context of the broader trial findings and reported trial detail.

Key Points about the Validity of the Findings

While we appreciate the importance of the types of issues addressed in the *Findings* report regarding whether the interventions were "really shown to be effective," we highlight how the *Findings* application of these issues as they relate to our article is misleading in a number of respects. There is a related statement that the findings should be reviewed "through a stronger lens." Key points for further reader consideration about what should constitute a "stronger lens" are delineated below.

Consider *All* Findings Relevant to PROSPER Efficacy. Of particular concern is the *Findings'* narrow focus on outcomes for the nine 12th grade current and frequency of use measures to support its conclusion that "the programme would have to have been declared largely a failure, or at least, not proven to be a success."

An examination of the PROSPER effects reported in the reviewed article, along with prior and subsequent outcome articles, shows a consistent pattern of favorable intervention-control differences and a complete absence of unfavorable differences across three developmental stages. *We think the sum total of substance misuse-related outcomes should be considered in a discussion of the intervention's overall success or public health impact.*

For example, the *Findings* narrow focus fails to factor reported effects on ***lifetime*** illicit substance use during both the 11th and 12th grades, despite evidence that delaying substance misuse onset reduces the future likelihood and severity of misuse (US Department of Health and Human Services [HHS], Office of the Surgeon General, 2016).

In addition, although the *Findings* commentary makes explicit reference to a total of 18 measured outcomes at each of the 11th and 12th grade time points, the *Findings*' review focuses only on nine current/frequency of use measures at 12th grade, with only one being judged "likely to remain statistically significant," using two-tailed tests and an unspecified correction for multiple tests. However, with a focus on the full range of outcomes, even with an application of a two-tailed significance level criterion (doubling all the reported *p*-values), 8 of 18 outcome measures were significant for 11th graders, and 7 of 18 for 12th graders, as well as were 8 of 18 growth trajectory (6th to 12th grade) outcomes. As concerns corrections for multiple tests, there is some debate about both need for corrections and the choice of methods when a correction is applied (Holm, 1979; Hochberg, 1988; McLaughlin & Sainari, 2014; Rothman, 1990). Even when a Bonferroni-Hochberg method is applied, only one more measure at each time point would not be counted as significant.

More importantly, the *Findings'* "largely a failure" conclusion ignores the breadth of published findings from the PROSPER trial reporting positive effects on a range of substance-related outcomes across three developmental stages. This includes a report on subsequent young adult (age 19) outcomes; in that case, 11 of 21 measures attained two-tailed significance (Spoth et al., 2017a). It also neglects published articles reporting social network effects and gene x intervention-related effects (e.g., Cleveland et al., 2015; Osgood et al., 2013; Ruilson, Gest & Osgood, 2015; Vanderberg et al., 2016).

Finally, the choice for presentation of the "largely a failure" point is noteworthy. The stand-alone, enlarged text box reads: "On the critical issue of how pupils ended up, ***only 1 of 9 measures produced a significant finding***" (our emphasis). This captures some of the wording from the "largely a failure" sentence in the text, adding the point about "1 of 9…" as a salient, definitive statement, even though arriving at the "1 in 9" reduction was a result of debatable choices, as discussed above.

Misleading Representation of School District and Pupil Dropout. Regarding the comment that the analyses for our study were not "intent to treat," the basis for this *Findings* concern was that "every school district and pupil randomly allocated" needed to be included in the analysis. Related to this issue, there was a specific point about how the loss and replacement of two

study sites at the beginning of the trial might explain the reported findings. Unfortunately, the loss of subjects and/or sites is a "fact of life" for community-based intervention outcome studies. Notably, as referenced in the *Findings* report, a conservative re-analysis was conducted for the Coalition for Evidence-Based Policy. The *Findings* summary of this re-analysis was misleading. In this re-analysis, two replacement sites in the invention group were excluded from the analysis along with the two control communities with the *highest* overall rates of substance use at 6.5-year follow-up. The resulting effects were only marginally smaller than across the full sample. This is especially noteworthy, since the re-analysis both reduced the degrees of freedom available for testing the intervention effects (which is based on the number of sites, not the number of individuals, in the study), as well as maximally reduced the observed intervention-control group differences at the site level. Further, the statement "Before the baseline measures had been collected, two of the school districts dropped out" is technically incorrect. The larger point is that more careful review of the detail of how this issue was addressed (e.g., in the linked references to the Coalition and Evidence-Based Policy Review) is indicated since it is highly unlikely this issue of site replacement affected the basic findings of the study.

As concerns pupil dropout, the *Findings* commentary stated that there was a failure to estimate results for those pupils that dropped out of the study before the 12th grade analyses were conducted. First, the analyzed sample included participants who completed surveys at 3 or more of the 8 data collection points, with full information maximum likelihood (FIML) used to address missing data within this somewhat reduced sample. (The selection of at least 3 out of 8 data points was made to more accurately describe the growth trajectories that were part of our analysis.) Thus, it is misleading to state that was "a failure to estimate results for the third of the intended sample of pupils not re-assessed at the 12th grade follow-up." As stated, FIML accounted for the missing data and it is misleading to suggest that "no attempt was made to reduce this risk by estimating on the basis of known data what the missing questionnaire responses would have been."

Second, we report how we addressed evaluation of differential drop-out or attrition. That is, we examined differential attrition on both sociodemographic measures (e.g., gender, age, race, school lunch status) and all substance misuse outcomes. It was assessed by examining whether the two-way interaction of Condition $\times$ Pretest score on the outcome variables predicted drop-out at each wave, and no significant interactions were found. Thus, it is highly unlikely that differential attrition played a role in the study findings.

<u>Inaccuracy of Representation of Researcher Allegiance Issue</u>. With regard to the reference to the lead author as "Creator, Advocate…and Examiner," the *Findings* commentary incorrectly states that "the lead author was the developer of the only intervention chosen by all the PROSPER teams." Although, as described in the program manual, development of a prior version of the program was funded by an NIH grant led by the lead author of the outcome article in question, none of the article authors were authors of the program. The lead author of the program was employed by the Extension and Outreach System (an administratively

separate unit of the University); the Extension Outreach System owns and administers the program.

In this regard, while we agree that it is important to be very cognizant of the threat of "researcher allegiance"—both as consumers *and* producers of research—it is also important to recognize that the substantial body of published findings from the PROSPER program of research has been vetted and shaped by numerous peer reviewers and editors who have no stake whatsoever in the outcome of this particular trial. This vetting included consideration of our analytic methods, measures, significance testing, sample composition, and conclusions.

Need to Consider A Broader Perspective on Public Health Relevance. Concerning our preceding argument to consider *all* relevant findings, it should also be noted that small differences in prevalence rates for serious substances, such as methamphetamines are, in fact, of public health relevance. In addition, delays in age of substance initiation are also of public health relevance, even if later prevalence rates are not affected, in part because of age-related substance effects on brain development. For example, the finding that there is a 42% reduction in new methamphetamine users by 10th grade (Spoth et al., 2011) is of public health significance and could lead to substantial cost savings, considering the estimated average cost to society in the case of a meth user ($33,606) and in the case of a meth addict ($74,000, see Nicosia, Pacula, Kilmer, Lundberg, & Chiesa, 2009).

Misleading Statements about Risk-related Analyses. The *Findings* commentary suggests that there was no pre-data analysis plan for our risk-related moderation. This is incorrect. The original PROSPER proposal that was approved by a Study Section at the National Institutes of Health proposed these analyses as a primary research aim. There also was reference in the commentary to not knowing how many of these pupils there were, implying that proportion of higher-risk students might have been small, yielding unreliable results. Higher-risk student representation in the sample was nearly 30%.

Need for a Broader Perspective on One-tailed Tests and Probability of Negative Findings. As indicated in the PROSPER outcome article, specific F- and *p*-values were included so that two-tailed significance levels could be easily calculated; there were no negative effects in evidence. As noted, all intervention effects were in the expected direction at earlier waves in this project, and prior evidence of program effectiveness was a criterion for program inclusion. In balance, numerous reviewers at top-tier journals have found reporting *p*-values to be appropriate; further, methodological literature-based reasoning for emphasis on one-tailed tests was accepted by reviewers and editors of all earlier reports in which they were applied. In this vein, methodologists from a number of fields have highlighted valid points to be made on both sides of the debate; some in the context of the need to balance scientific rigor with the need for practical knowledge (e.g., see Cho & Abe, 2013; Good, 1992; Graham, 2008; Lakens, 2014).

Need to Consider Context of Representation of Pre-trial Study Plans. We fully support the move towards the registration of pre-trial study plans; we have done so in a PROSPER replication recently undertaken. Such registrations, however, were not common practice when the

reported study was begun in 2002. The original PROSPER proposal was reviewed and approved by a Study Section at National Institutes of Health. The proposal specified the measures and analyses that would be conducted; changes in that plan were due to developmental changes in the sample across study funding cycles (at the time this study was begun, we could not have reasonably anticipated that the study would follow youth into later adolescence, epidemiology-related changes that would lead to changes in measurement (e.g., the emergence of prescription drug abuse as a significant public health problem), and advances in analytic techniques readily available to researchers.

Key Points of Relevance to Replicability Issue

Generalizability Limitations Already Noted. In the context of raising questions about whether the tested system "would work elsewhere," we appreciated the acknowledgement that generalizability limitations are stated on the PROSPER outcome article, explicitly indicating the additional study regarding the practical viability of the PROSPER delivery system in differing types of community settings.

Need to Consider a Broader Range of "Real World" Expectations for Generalizability. It is striking that the "real world" issue about whether the intervention would work elsewhere is raised, while failing to note other real world issues of relevance. Prior PROSPER reports clearly highlight how generalizability pertains to "ready" communities both willing and able to implement the model. As described in prior PROSPER reports, its viability required an interested school district in a location where a Cooperative Extension educator was available. We have never suggested that findings are "…a guide to what would happen in places like London and New York." Articles describing the PROSPER model explicitly state that it was designed for the types of school districts and communities actually enrolled in the study. In that connection, it is noteworthy that there is an estimated pool of around 6,000 communities or towns with a population up to 50,000 in the US alone; all of those are located in states served by land grant universities with Extension Outreach systems.

Misrepresentation of Protocol Guiding School District Recruitment. It is also inaccurate to state that the 15 schools in the original list that were left out on "undocumented grounds." The study protocol (approved by the NIH funding agency) established the number of intervention and control condition schools that would be necessary to avoid Type 2 errors. After that number was reached, additional schools were not contacted; more than 28 schools would have been financially prohibitive.

Concluding Comments on Broader Issues of Relevance

Again, in general, we commend *Findings* for raising and discussing important issues relevant to the interpretation of reported findings from the PROSPER and other prevention trials. That said, we encourage a more balanced and comprehensive consideration of the findings from the reviewed article in the context of the broader trial findings and reported trial detail.

First, we would like to draw the readers' attention to prior commentaries addressing similar validity issues, including one on the validity of PROSPER findings (Rulison, Feinberg, Gest, & Osgood, 2016; Spoth, Trudeau, Redmond, & Shin, 2008; Spoth et al., 2017b; Spoth, Trudeau, Redmond, & Shin, 2009). We encourage readers to review these relevant commentaries. In particular, we would like to highlight earlier responses concerning the point in the current *Findings* commentary about issues with the evidence on one of the programs on the PROSPER menu (the Strengthening Families Program for Parents and Youth 10-14). Notably, studies of the SFP 10-14 program that are cited as raising a question about its efficacy have many differences from the studies conducted by our team, including the fact that none followed participants for more than 3 years post-implementation. For example, some of the studies were conducted with younger students among whom there are very low prevalence rates; our studies indicate that differences between intervention and control groups emerge later, when given types of substance use are more normative. Especially to the point, as noted previously, the summary of the findings from other studies ignores the full range differences from the original trial—in designs, samples, intervention adaptations, and country and cultural contexts—factors that should be considered in a balanced discussion of the generalizability of findings and the current efficacy of the program with rural US populations and elsewhere. We would encourage readers to search out the original articles and commentaries for a more balanced and complete view. Similar to the first key validity point about considering *all* of the relevant data, in balance, there are additional articles providing results from both of the referenced earlier research projects (e.g., into young adulthood) that were not cited in the commentary *or* linked documents.

Second, we agree that it is appropriate to consider application "beyond a reasonable doubt" criteria to scientific evaluations. Our concern is the lack of specificity about how these criteria are operationally defined and applied, especially in the case of an article publishing findings from a longitudinal prevention trial originating in 2002 and conducted through the present, over a period of 15 years. In this vein, our view is that there is a need to apply *appropriate* rigor to the reporting and interpretation of findings – balancing "beyond a reasonable doubt" with "the preponderance of evidence" to arrive at the most reasonable conclusion. Although we recognize that opinions on how to operationalize such criteria may differ, the point of the research endeavor should be to advance the science and produce useful knowledge.

Finally, it is worth noting that the lead PROSPER investigators recently wrote a commentary that summarized guidelines for *constructive criticism* (Spoth et al., 2017b). A number of these guidelines are of relevance in the case of the present commentary, including careful attention to methodological detail across multiple reports from large programs of research.

A copy of the original *Drug & Alcohol Findings* review article to which this commentary responds is available upon request.

References

Cho, H-C, & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research, 66,* 1261-1266.

Cleveland, H. H., Schlomer, Vandenbergh, D. J., Feinberg, M., Greenberg, M., Spoth, R., Redmond, C., Shriver, M. D., Zaidi, A. A., & Hair, K. L. (2015). The conditioning of intervention effects on early adolescent alcohol use by maternal involvement and DRD4 and 5-HTTLPR genetic variants. *Development and Psychopathology, 27*(1), 51-67. Cambridge University Press: http://dx.doi.org/10.1017/S0954579414001291

Good, I. J. (1992). The Bayes/Non-Bayes Compromise: A Brief Review. *Journal of the American Statistical Association*, *87*(419), 597. http://doi.org/10.2307/2290192

Graham, K. (2008). Fiddling while Rome burns: Balancing rigour with the need for practical knowledge. A Commentary. *Addiction*, 103, 414-415.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometricka,* 75, 800–802.

Holm, S. A. (1979). Simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. http://doi.org/10.1002/ejsp.2023

McLaughlin, M. J. & Sainari, K. L. (2014). Bonferroni, Holm, and Hochberg corrections: Fun names, serious changes to *p* value. *The American Academy of Physical Medicine and Rehabilitation,* 6(6), 544-546.

Nicosia, N., Pacula, R. L., Kilmer, B., Lundberg, R., & Chiesa, J. (2009). The economic cost of methamphetamine use in the United States, 2005. Retrieved 9-25-17, at https://www.rand.org/pubs/monographs/MG829.html

Osgood, D. W., Feinberg, M. E., Gest, S. D., Moody, J., Ragan, D. T., Spoth, R., Greenberg, M. & Redmond, C. (2013). Effects of PROSPER on the influence potential of prosocial versus antisocial youth in adolescent friendship networks. *Journal of Adolescent Health, 53*(2), 174-179. http://dx.doi.org/10.1016/j.jadohealth.2013.02.013

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1 (1), 43-46.

Rulison, K., Feinberg, M., Gest, S. D., & Osgood, W. (2016) Response to Forman et al., "Comment on Rulison et al. (2015). Diffusion of Intervention Effects." Developmental Psychology, 58(6), 693. http://www.jahonline.org/article/S1054-139X(16)30005-2/pdf

Rulison, K. L., Gest, S. D., & Osgood, D. W. (2015). Adolescent peer networks and the potential for the diffusion of intervention effects. *Prevention Science, 16,* 133-144. http://dx.doi.org/10.1007/s11121-014-0465-3

Spoth, R., Redmond, C., Clair, S., Shin, C., Greenberg, M. & Feinberg, M. (2011). Preventing substance misuse through community-university partnerships: Randomized controlled trial outcomes 4.5 years past baseline. *American Journal of Preventive Medicine, 40*(4), 440-447. http://dx.doi.org/10.1016/j.amepre.2010.12.012

Spoth, R., Redmond, C., Shin, C., Greenberg, M., Feinberg, M., Trudeau, L. (2017a). PROSPER delivery of universal preventive interventions with young adolescents: Long-term effects on emerging adult substance misuse and associated risk behaviors. *Psychological Medicine, 47(13), 2246-2259.* http://dx.doi.org/10.1017/S0033291717000691

Spoth, R., Redmond, C., Shin, C., Greenberg, M., Feinberg, M., Trudeau, L. (2017b). Sources of Bias in Gorman Critique of Bias: Again a Need for More Reasonable, Valid Conclusions with True Dialogue. *Psychological Medicine EPub.*

Spoth, R., Trudeau, L., Redmond, C., & Shin, C. (2008) Finding a path to more reasonable conclusions about prevention—Response to Midford. *Addiction, 103(7),* 1169-1173. Letter to editor http://dx.doi.org/10.1111/j.1360-0443.2008.02270.x

Spoth, R., Trudeau, L., Redmond, C., & Shin, C. (2009). Further clear examples of the need for more reasonable conclusions and critiques about prevention. *Addiction, 104*, 154-155 http://dx.doi.org/10.1111/j.1360-0443.2008.02460.x

US Department of Health and Human Services [HHS], Office of the Surgeon General, (2016). *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health.* Washington, DC: HHS, November 2016. Retrieved 9-25-17, at https://addiction.surgeongeneral.gov/surgeon-generals-report.pdf

Vandenbergh, D. J., Schlomer, G. L., Cleveland, H. H., Schink, A. E., Hair, K. L., Feinberg, M. E., Neiderhiser, J. M., Greenberg, M. T., Spoth, R. L., Redmond, C. (2016). An adolescent substance prevention model blocks the effect of CHRNA5 genotype on smoking during high school. *Nicotine & Tobacco Research*, 18(2), 212-220.

Response to comments from Dr. Ashton concerning "Commentary on Research Analysis from PROSPER community-university partnership delivery system effects on substance misuse through 6½ years past baseline from a cluster randomized controlled intervention trial."

(Submitted to *Drug & Alcohol Findings* on 12-4-17)

Spoth R., Redmond C., Shin C., Greenberg M., Feinberg M., Schainker L.

Overview of Dr. Ashton's Responses to our Earlier Commentary: We appreciated the time you took to further evaluate more of our PROSPER outcome reports, particularly those with additional data on PROSPER outcomes post high school, your thorough response to our comments below, and the related editing of your original review. That said, we remain concerned about how you return to very similar conclusions on most of the key issues, and especially how there are many instances of selective attention to the literature of relevance, particularly concerning our overriding issue of a balanced representation of all of the relevant PROSPER findings.

The following comments on your responses provide many concrete examples that collectively raise considerable concerns about a truly balanced review of all relevant PROSPER findings. In this connection, there are cases where you admit you do not have direct evidence of bias but clearly suggest it could very well be the case (e.g., the repeated references regarding "publically available" pre-trial analysis plans, despite our earlier description of pre-trial detail in our NIH-funded proposal). There are other cases where you argue for selective attention to subsets of findings (e.g., ostensibly, to be responsive to reader interest) that nonetheless are central to your global assessments of "prevention impacts." There are yet other instances where your choice of words (e.g., retention of the descriptor "creator" of the SFP 10-14) seems inconsistent with factual information about its development. Finally, in the context of explicit reference to our bias, you fail to mention clearly-relevant analytic detail provided in our reports (e.g., the rationale and oversampling methodological detail in the age 19 outcome paper review that would run counter to bias).

Briefly put, if there only were infrequent indicators of selective attention or subjectivity, we would have no concern about bias in your critique, all things considered. Although offset somewhat by our positive view of your effort to provide a quality critique, it is, nonetheless, somewhat akin to the major point in the Gorman rejoinder that you characterized as "rather persuasive." Although we consider your critique to be much more thorough, factually sound and better reasoned than that of Gorman, we fear that the number of cases of subjective judgement and selective attention to relevant information (not clearly represented or justified as selective or subjective), warrants more general, careful consideration of places where your own views might be biasing your presentation. All co-authors were impressed in a similar way. I hope this impression becomes clear through highlighted responses below.

Dr. Ashton's comments

1. The Main findings section intended to reflect the paper's findings alludes to all the results presented in tables 1 and 2 of the featured report including lifetime use, and we also pulled in results from

supplementary table 1. This seems a reasonably comprehensive account of the findings as presented. With no indication of what for the authors are the main outcome measures, in the commentary we chose to highlight those we considered of greatest significance for our readers.

We explicitly indicated that our primary concern about your review was its narrow focus on outcomes for nine 12th grade current and frequency of use measures in the context of claims about the preventive impacts of PROSPER outcomes overall.

We acknowledge and appreciate the effort to include reference to some of the additional PROSPER outcome articles we noted, particularly the one reporting on young adult outcomes. That said, we conclude that you have not fully considered individual reports of outcomes (e.g., our age 19, young adult outcomes paper, see below) and do not see evidence of close scrutiny of the sum total of the outcomes across multiple reports, as noted under our comments to this response and others following.

First, we would like to note your statement that "…we chose to highlight those we considered of greatest significance for our readers." This illustrates a judgement call, and represents selective reporting by our view, especially considering that how you represent those self-selected findings is part of an argument that impugns the sum total of PROSPER prevention impacts. Since the primary article that is the focus of your critique provides Table 1 and 2 in the manuscript, at least all the outcomes in both tables (not even considering the supplementary online table) should be addressed equally.

Second, you support your decision to focus on the selected 12th grade outcomes by diminishing the importance of preventive intervention effects during adolescence on subsequent young adult outcomes, when there is considerable empirical evidence to the contrary for community-based preventive interventions. For example, see the long-term data from the Midwestern Prevention Project; they show lower rates of adult amphetamine use and lower mental health utilization (citations at end). In addition, longitudinal findings of Communities that Care also show positive young adult outcomes as a result of preventive interventions during adolescence (citation at end).

Third, longitudinal prevention trials through 12th grade often report more attrition than in earlier grades in high school, and lower completion rates of surveys administered in schools. This is another reason the choice of a selected subset of 12th grade finding as the final arbiter of impact is problematic.

Fourth, there remains a technically incorrect statement: It is: "The largest relative reduction in the proportion of pupils was in respect of past-year methamphetamine use at the 12th grade, a reduction from about 4% to about 3%." The finding is 4% vs. 2% for the higher-risk group.

2. We cannot simply assume that delaying onset of use through PROSPER reduces later substance use problems. Findings that later onset is associated with fewer problems may reflect background features of the lives of the people concerned. When delay has been engineered via a prevention programme, the same association will not necessarily be found.

You say "simply assume… may reflect background features of the lives of the people concerned." We are not entirely clear about how background features can explain away our findings, or other ones (see

response above) in the literature. Your argument dismisses the relevant literature regarding age of onset and the likelihood of substance disorders (not to mention our finding of reduced drug-related problems at age 19), in favour of an unsubstantiated assumption that no such relationship will exist for prevention programme participants. This dismissal also is not consistent with the scientific literature directed toward the economic benefits of prevention (see the Washington State Institute for Public Policy on economic benefit of prevention, http://www.wsipp.wa.gov/BenefitCost).

3. But if this is the case, one would hope the effects remained visible at the latest follow-up, in this paper, in the 12th grade. If by then they are not statistically significant, that does not bode well for effects in later years. Having described all the findings in the Main findings section, it seemed reasonable to focus on the final in-school follow-up in the Commentary. It is not much comfort if at earlier years there were impacts and the growth curves were less steep, if how the pupils ended up does not significantly differ. But have now in the Commentary prefaced the relevant section (Focus on the final in-school follow-up) with a reference to the overall pattern of the findings, changed the heading from 'Reviewing the findings through a stronger lens' to 'Focus on the final in-school follow-up', and further explained the decision to focus on the 12th grade. From the start we had said early in the Commentary that "On the balance of probabilities, it is likely that there were preventive impacts."

First, although we appreciate the addition of the preface, it only acknowledges and does not address the fundamental issues with selective reporting that pave the way to reaching a global conclusion about prevention impacts. You selectively focus on "the latest follow-up in this paper" and the "final in-school follow-up." It simply is not clear why would readers care more about 12th grade findings than findings at age 19.

Second, again, you are assuming that delayed initiation is irrelevant, and that growth trajectories across one developmental phase have no bearing on behaviours during a subsequent phase. In this connection, note the findings cited earlier (e.g., long-term data from the Midwestern Prevention Project showing lower rates of adult amphetamine use and lower mental health utilization).

In this vein, you are dismissing some fundamental "givens" about growth curve analyses across developmental stages. PROSPER entailed collecting multiple waves of data before, during, and after high school. The developmental stages of students need to be considered in this context. Growth curve patterns of substance misuse outcomes frequently have linear relationships within developmental stages but S-shaped growth patterns across stages. Focusing on steepness of the slope within a stage is only a part of the overall growth pattern. For example, regardless of the steepness of the within-stage slope in lifetime alcohol prevalence rates, the proportion of lifetime alcohol use will approach 1 at some point as a lifetime S-shaped growth pattern.

Finally, you say "However, criteria closer to 'beyond reasonable doubt' are commonly applied in scientific evaluations. Then you proceed to cite untested (and unclaimed) generalizability to a different population in a completely different context as an application of "beyond reasonable doubt" criteria.

4. Agree with you that the term 'failure' is inappropriate and it has been removed throughout.

We appreciate the acknowledgement and the related editing.

5. The '1 in 9' quote has been deleted – too complex I agree to be captured in a quote – and the main text softened somewhat.

Agreed!

6. Have as you suggest added in findings from the later report when participants aged 19 and also the first outcomes report. This is the section which adds new material: http://findings.org.uk/PHP/dl.php?f=Spoth_R_25.txt#other

Again, we appreciate the added summary of findings from the age 19 outcome paper and the initial reference to how the overall pattern of findings did favor PROSPER. Our primary concern here is the specific manner in which you dismiss that pattern of findings.

To the larger point about selective reporting and how it connects to critique bias, when considered as a whole, the new text clearly suggests we "cherry-picked" the age 19 results. The summary also fails to consider outcomes from indices of illicit drug use, for example, that have been relatively consistently significant across PROSPER reports. In addition, you initially only refer in passing to the key drug-related problems outcome obviously germane to the consistent illicit drug use findings. Furthermore, your review uses the findings from the 1.5 year follow up assessment to suggest that short-term follow-ups are not the reason why other studies "have not found components of the PROSPER model effective." Overall, it appears that rather than objectively evaluating findings from those additional reports in the context of all of the reported findings from PROSPER you are interpreting findings with a slant toward your pre-existing conclusions.

The fact that you did not consider other outcomes noted in our earlier response (e.g., reports on Gene x Environment Interaction effects) contributes to our concerns. More specific concerns about your representation of findings from our age 19 outcome paper are as follows.

First, you fail to note that our sub-sample sustained randomization and was randomly selected, or that we evaluated how it compares to the overall sample.

Second, your statement in the last sentence of the sixth paragraph of the supplemental text ("… in the process biasing the analyzed sample in favor of PROSPER…") again suggests a biased reviewer perspective, given the complete lack of reference to the fact that the oversampling (randomly selected) was to ensure the validity of appropriately examining outcomes for the higher-risk subsample, its explicitly expressed purpose. Plus, increasing the proportion by 8% is not that substantial, in any case.

Third, despite our earlier response concerning issues with assumptions that Bonferroni should be a "gold standard" (not to mention our demonstrating how reasonable corrections for multiple tests failed to explain away positive outcomes on the high-school outcomes articles) you fail to represent those issues.

Fourth, there again is reference to attributions about specific components of PROSPER when we

highlighted the considerable constraints on such attributions (given the PROSPER experimental design) in our prior response.

Fifth, considering the extent to which analytic detail provided suggesting the possibility, you do not factor the possibility that underpowered analyses could have contributed to the patterns of findings.

Finally, regarding the last sentence in the last paragraph of the supplement, you could have checked on the details from our sample tracking before indicating the issues you note.

7. Thanks for pointing out that the statement "Before the baseline measures had been collected, two of the school districts dropped out" is technically incorrect. Now replaced with "During the first year of the trial".

Thank you.

8. You say our summary of the analysis conducted for the Coalition for Evidence-Based Policy "was misleading", pointing out that "two replacement sites in the invention group were excluded from the analysis along with the two control communities with the highest overall rates of substance use at 6.5-year follow-up. The resulting effects were only marginally smaller than across the full sample. This is especially noteworthy, since the re-analysis both reduced the degrees of freedom available for testing the intervention effects (which is based on the number of sites, not the number of individuals, in the study), as well as maximally reduced the observed intervention-control group differences at the site level."

What we said was that "The Coalition was concerned that the drop-out of two PROSPER school districts undermined the randomisation intended to assure a level playing field, so instead of all 28 districts, they relied on a re-analysis done by the study's researchers which 'omitted the two replacement communities in the PROSPER group along with the two control communities with the highest overall rates of substance use at 6.5-year follow-up'. The resulting effects were marginally smaller than across the full sample."

I admit I can't see how this was misleading. Also, regardless of the findings of this re-analysis, the exclusion of the two drop-out communities is a departure from the claimed intent-to-treat analysis. However, I can see that more could have been said, so now reads: "Excluding the 'worst' of the comparison districts would have tilted the analysis against PROSPER, yet the estimated effects were only marginally smaller than across the full sample. Nevertheless, this analysis still contravened intent-to-treat principles, and it remains possible that the two of the 14 districts allocated to PROSPER which dropped out would have registered worst outcomes than even the worst of the comparison communities." Elsewhere have added "However, the re-analyses which were done suggest that the impacts of this departure from [intent-to-treat] standard were probably small."

Again, although we appreciate the rewording, our concern remains about how you represent the analyses.

First, although the intervention district drop outs could have had the worst outcomes, they also could

have had the best outcomes. It simply is unknowable. The analyses of the possible effects of the district replacements were rigorous by most standards and they suggested minimal impact.

Second, it simply is true that the sample for any study might not be representative. We can only perform the types of analyses we did to gain confidence it is representative and the results are valid.

Third, we believe the framing for the intent-to-treat discussion ignores the key point of relevance to the reported findings (and the key element of an intent-to-treat analysis more generally). That is, participants assigned to the intervention group were included in the analysis regardless of whether or not they received any intervention; everyone that provided the necessary data was included, and no one was excluded for intervention non-compliance. Participants frequently change their minds after initially agreeing to be involved in a study, with the result that there are no data to include in the analysis. Communities, as the unit of analysis here, can do the same. Because the intent-to-treat principle was applied to the analysis of all participant data collected, we're unclear on the point of your intent-to-treat discussion.

9. You say, "As concerns pupil dropout, the Findings commentary stated that there was a failure to estimate results for those pupils that dropped out of the study before the 12th grade analyses were conducted. First, the analysed sample included participants who completed surveys at 3 or more of the 8 data collection points, with full information maximum likelihood (FIML) used to address missing data within this somewhat reduced sample. (The selection of at least 3 out of 8 data points was made to more accurately describe the growth trajectories that were part of our analysis.) Thus, it is misleading to state that was 'a failure to estimate results for the third of the intended sample of pupils not re-assessed at the 12th grade follow-up.' As stated, FIML accounted for the missing data and it is misleading to suggest that 'no attempt was made to reduce this risk by estimating on the basis of known data what the missing questionnaire responses would have been.'"

   I take your point here, but my understanding is that this FIML analysis applied only to the growth trajectory analyses ("To improve validity of the growth trajectory estimates, students were included if they completed surveys at three or more of the eight data collection points. Missing data were handled using full-information maximum likelihood estimation.") and our focus in the commentary was on the final follow-up results, which I assumed was understood by readers. This is why we said (emphasis added), "*Except for the trend over the whole follow-up*, no attempt was made to reduce this risk by estimating on the basis of known data what the missing questionnaire responses would have been."

This is incorrect.

FIML was used to account for missingness in both growth trajectory and point-at-time models. FIML is generally considered to be comparable or superior to imputation techniques (including multiple imputation) for yielding unbiased parameter estimates when there are missing data.

   I have now made this distinction clearer still: "Another shortfall from this standard was that when assessments were made of how pupils ended up at the 12th grade follow-up, results for the third of

the pupils not re-assessed (27% of the baseline sample) were not estimated for inclusion in the outcomes analysis, a loss which reviewers said might lead to inaccurate estimates of PROSPER's effects. This is likely to have been particularly important when so few youngsters engaged in the substance use pattern being assessed that a missing few could have easily have tipped the balance – the case, for example, in respect of the most substantial 12th-grade findings, which related to methamphetamine use. Except for the trend over the whole follow-up, no attempt was made to reduce this risk by estimating on the basis of known data what the missing questionnaire responses would have been."

If this remains incorrect I'd be pleased to be corrected.

Please see our comments above, our comments in our original response, and our comments in the Gorman critique you cite.

While it is reasonable to be particularly cautious when considering results concerning very low frequency behaviors, our capacity to examine such behaviors in the PROSPER study is facilitated by our large sample size (by design).

10. You say, "Second, we report how we addressed evaluation of differential drop-out or attrition. That is, we examined differential attrition on both sociodemographic measures (e.g., gender, age, race, school lunch status) and all substance misuse outcomes. It was assessed by examining whether the two-way interaction of Condition x Pretest score on the outcome variables predicted drop-out at each wave, and no significant interactions were found. Thus, it is highly unlikely that differential attrition played a role in the study findings."

    Yes, but the point being made was to question the claim of an intent-to-treat analysis, not to say whether this actually biased the results. However, now reads: "Checks found that on known variables there were no significant differences between the kinds of pupils not followed up in PROSPER versus districts, leading the researchers to argue that loss to follow-up was unlikely to have biased the findings. However, such checks cannot eliminate the possibility that, with small differences, a more complete follow-up would have produced different results."

Please see above comments on intent-to-treat analyses under item #8.

We do agree that results could have been different (that always is the case), but we question the reasoning here. Different samples will produce different results; that is not a basis for concluding that those results are likely to be less strong – in the absence of any rationale to contrary, they're just as likely to be stronger.

11. You say, "Although, as described in the program manual, development of a prior version of the program was funded by an NIH grant led by the lead author of the outcome article in question, none

of the article authors were authors of the program. The lead author of the program was employed by the Extension and Outreach System (an administratively separate unit of the University); the Extension Outreach System owns and administers the program."

Thanks for correcting this. Now says: "The lead author of the featured trial led the project which initially developed the only intervention chosen by all the PROSPER teams.

It still seems quite problematic to use the label "creator" in the title of the side box when that literally was not the case, as described earlier. Retaining that label will lead any reader not carefully attending to the detail relevant to the present study to falsely conclude the author(s) to be program creators.

Also, rather than "…which initially developed the only intervention chosen by all the PROSPER teams," it would it be better to say "… for which the family intervention selected by all PROSPER teams was developed" – note that we think mention of "family program" is important here, since the teams also selected a school program (not developed at ISU).

12. You say, "In this regard, while we agree that it is important to be very cognizant of the threat of 'researcher allegiance'—both as consumers and producers of research—it is also important to recognize that the substantial body of published findings from the PROSPER program of research has been vetted and shaped by numerous peer reviewers and editors who have no stake whatsoever in the outcome of this particular trial. This vetting included consideration of our analytic methods, measures, significance testing, sample composition, and conclusions."

    Yes, but it is also the case that some of those assessors rejected the one-tailed testing strategy adopted in the featured article, took into account adjustments for multiple tests not taken into account in the featured article, and saw the findings as vulnerable to bias due to attrition, a possibility not mentioned as a limitation in the featured article. And peer-review publication and accolades from organisations such as Blueprint do not guarantee methodological adequacy. A quite blatant example is Project STAR – see http://findings.org.uk/PHP/dl.php?f=Ashton_M_21.pdf

You probably realize that you are characterizing a large number of reviewers, given the number of reports on PROSPER outcomes (especially including those from "spin off " studies on social network outcomes, and GXE interaction effects primary substance misuse and conduct problem outcomes). Thus you are saying that there would be a very considerable number of reviewers who "…rejected the one-tailed testing strategy adopted in the featured article, took into account adjustments for multiple tests not taken into account in the featured article, and saw the findings as vulnerable to bias due to attrition…" without so indicating in their reviews and requiring us to address such. That certainly does seem unlikely on the surface of it and we wonder about the objective basis for that kind of statement. Also, absent seeing the detail about Project STAR we are not in a position to critically evaluate the degree to which it is methodologically inadequate. We suspend judgement until we critically evaluate the reviews. That said, after reading this response (including the language in reference to Project STAR like "quite blatant example"), we are more worried about a lack of a balanced perspective.

13. You say, “Concerning our preceding argument to consider all relevant findings, it should also be noted that small differences in prevalence rates for serious substances, such as methamphetamines are, in fact, of public health relevance. In addition, delays in age of substance initiation are also of public health relevance, even if later prevalence rates are not affected, in part because of age-related substance effects on brain development. For example, the finding that there is a 42% reduction in new methamphetamine users by 10th grade (Spoth et al., 2011) is of public health significance and could lead to substantial cost savings, considering the estimated average cost to society in the case of a meth user ($33,606) and in the case of a meth addict ($74,000, see Nicosia, Pacula, Kilmer, Lundberg, & Chiesa, 2009).”

In general, our article provides both results of statistical tests (including p values) and clinical significance findings (relative reduction rates). With this, the reader is provided with what they need to critically evaluate the findings overall.

> We allude to this in saying “A further important consideration is whether, even assuming they were real effects, the generally small gains achieved by PROSPER are worth the considerable investment it took to produce them.” It is rather a large step from small reductions in the spread and frequency (not demonstrated for methamphetamine) of use to fewer addicts, given the other and much deeper factors which generate addiction. And this is not a one-way street. There are studies associating early substance use with better prospects in later life, and others suggesting any link between early onset and later problems is due to other factors which affect both and not causal.

Your statement “…even assuming they were real…” is another instance of language that raises red flags for us about potential critique bias.

Regarding your statement “There are studies associating…,” could you provide examples? And with regard to “other and much deeper factors,” did you consider putative preventive intervention effect mechanisms that include positive effects on parent-child relationships, youth decision-making, refusal skills, and other factors associated with the development of substance misuse and other problem behavior. Also, did you consider our published effects on social capital? We are concerned here about whether there has been a balanced review of the literature.

> The most hopeful finding is the age-19 reduction in drug-related problems but it is unclear how to interpret the difference of from just over to just under 1 on a range which I guess is something like from 0 to 15. Talk of cost-savings down the line is hopeful but speculative I suggest. Still have added a further reference to this finding: “A further important consideration is whether, even assuming they were real effects, the generally small gains achieved by PROSPER are worth the considerable investment it took to produce them. On this score perhaps the most hopeful finding is a reduction in drug-related problems found when the former pupils were aged 19, but it seems the scale to measure this was an unvalidated adaptation of an existing scale, it is unclear how to interpret what seems a very small difference, and how representative the sample was by that stage too is unclear.”

You have chosen to selectively omit a description of that “unvalidated adaptation” – specifically, the adapted measures’ more refined approach of asking about alcohol and drugs separately, rather than

having them confounded. Choosing to cite this as a threat to validity, while ignoring the clear substantive rationale for our approach, is another example suggesting a biased perspective in the review.

Again, as suggested earlier, sample representativeness seems to us to be a straw-man argument in the absence of any supportive evidence.

Given the actual effect size, we are not sure why you label the drug-related problem result as "very" small.

14. Under the heading, "Misleading Statements about Risk-related Analyses", you say: "The Findings commentary suggests that there was no pre-data analysis plan for our risk-related moderation. This is incorrect. The original PROSPER proposal that was approved by a Study Section at the National Institutes of Health proposed these analyses as a primary research aim. There also was reference in the commentary to not knowing how many of these pupils there were, implying that proportion of higher-risk students might have been small, yielding unreliable results. Higher-risk student representation in the sample was nearly 30%."

    What we said seemed correct and not misleading given the information made available in the article. There was no public pre-data analysis plan - but have now clarified to "publicly available". I did search for the research proposal but was unable to find it. It is good to know that the high v. low risk assessment was planned in advance, but without seeing the PROSPER proposal you refer to, we still cannot know how that division was to be made. So it is still the case that, as we said (emphasis added), that: "The main methodological concern is whether *this way* of dividing the sample into higher and lower risk was planned before the data indicated that it might produce desirable results." Thanks for telling us that the higher-risk student representation in the sample was nearly 30%. Have now added this in.

Regarding your reference to "this way," we take your point. That said, please see our introductory comments about the repeated reference to pre-trial analysis plans, and those under item 16. Also, we consider early gateway substance initiation to be an obvious and substantively defensible basis for defining risk throughout adolescence.

15. Regarding your comments under "Need for a Broader Perspective on One-tailed Tests and Probability of Negative Findings". Note that the main aim in our commentary was to test the reasons you gave for one-tailed tests, not to argue that it was wrong to use these tests – just that the reasons given do not stand up – in our opinion. I looked at the articles you cited. To my they do not seem to justify one-tailed tests, at least not at the 0.05 level:

Importantly, no negative findings were concealed and readers can apply the significance level of their choosing – weighing the consequences of Type 1 versus Type 2 errors, and how they view "the preponderance of evidence."

And again, statements like the "…authors are trying to give an impression for the robustness…" even

though exact p-values are provided, raise a red flag. It is reasonable to expect that readers will understand what a p-value means for either one-tailed or two-tailed tests. The statement suggests an inclination to take readers in an inappropriate direction, by our view.

Finally, the sentence including the citations refers to the "…the context of the need to balance scientific rigor with the need for practical knowledge." The citations are intended to support this statement as a whole.

Cho & Abe, 2013
Does argue for one-tailed tests but also says, "For correct directional decisions, we have noted that the two-tailed p values should be halved. Otherwise, researchers are likely to draw inaccurate or mistaken empirical conclusions at a given level of significance? (e.g., 'p=0.08: p>0.05' for inexact two-tailed testing, however, 'p=0.04: pb0.05' for exact one-tailed testing)." A useful point I have incorporated in the commentary.

Good, 1992
In my ignorance of the finer points of statistical analysis, I could not see the relevance to point at issue.

Graham, 2008
Is not an argument for one-tailed testing but for relaxing p values.

Lakens, 2014
Is not an argument for one-tailed testing but for interim hypothesis testing to avoid unnecessarily large samples.

16. Your comments under "Need to Consider Context of Representation of Pre-trial Study Plans" explains why there was no registration of pre-trial study plans and that is perfectly understandable – but our comment was not just about registration: "Pre-selecting a primary yardstick of success is considered critical to clinical trials." The citation given in support dates back to 1999. Selecting and publishing the selection of a primary yardstick does not require registration. Have made that clearer now.

This suggests all research findings in all fields are dismissed unless there is a published pre-trial study plan.

We are especially concerned about the frequent reference to pre-trial study plans, stated in connection to many points (across pages 6-10 in your original) especially in light of the historical context and factors detailed in our original response (e.g., our proposal reviewed by NIH). Across the multiple references to this issue, you are inferring or suggesting bias, absent evidence of such. Our view is that it would be more balanced to acknowledge the contextual factors and be more careful about your messaging.

17. You say, "The proposal specified the measures and analyses that would be conducted; changes in that plan were due to developmental changes in the sample across study funding cycles (at the time this study was begun, we could not have reasonably anticipated that the study would follow youth

into later adolescence, epidemiology-related changes that would lead to changes in measurement (e.g., the emergence of prescription drug abuse as a significant public health problem), and advances in analytic techniques readily available to researchers."

Yes accepted and now included in the text - but this does not seem to account for some of the changes in measurement (e.g., between age 18 and 19), and it would have remained possible to specify primary measures at each new analysis but before the data had been collected.

18. You say, "It is striking that the 'real world' issue about whether the intervention would work elsewhere is raised, while failing to note other real world issues of relevance. Prior PROSPER reports clearly highlight how generalizability pertains to 'ready' communities both willing and able to implement the model. As described in prior PROSPER reports, its viability required an interested school district in a location where a Cooperative Extension educator was available. We have never suggested that findings are '…a guide to what would happen in places like London and New York.' Articles describing the PROSPER model explicitly state that it was designed for the types of school districts and communities actually enrolled in the study. In that connection, it is noteworthy that there is an estimated pool of around 6,000 communities or towns with a population up to 50,000 in the US alone; all of those are located in states served by land grant universities with Extension Outreach systems."

    We did not suggest that you or the article "suggested that findings are 'a guide to what would happen in places like London and New York.'" We simply made that point, which in a UK context is important given the concentration of the population (and to a degree of drug problems) in large cities. But have now clarified: "How a community-based programme would be engaged with and implemented in these perhaps isolated, small, and homogenous communities, is not necessarily a guide (as the researchers acknowledge) to what would happen in places like London and New York."

We would prefer that you simply highlight that we have US findings based on populations acknowledged to be dissimilar and to carry generalizability constraints (e.g., rural US). It would be much more preferable than your making the case that the US research is flawed on this count (plus otherwise biased).

Take your point about PROSPER being designed for the communities you specify. Importantly it means that uncertain generalisability is not a criticism of the research but simply a fact. We should have made that clearer have now done so: "It is important to understand that the PROSPER system was developed for small rural communities with the required infrastructure, willingness and capacity. "Results are primarily expected to generalize to the type and size of communities selected for this study," acknowledged the researchers. Their reports highlight the importance of local school and university extension system personnel willing to engage in a collaborative effort to prevent substance use problems in their communities. Support from state-level university researchers and a land grant university's cooperative extension system mean the local teams are not left on their own, but still they have to do the bulk of the work. Inevitably this genesis is likely to limit the applicability of the PROSPER system."

This helps.

19. You say, “It is also inaccurate to state that the 15 schools in the original list that were left out on “undocumented grounds.” The study protocol (approved by the NIH funding agency) established the number of intervention and control condition schools that would be necessary to avoid Type 2 errors. After that number was reached, additional schools were not contacted; more than 28 schools would have been financially prohibitive.”

    Yes, understood. But the diagram on p. 192 of the featured report starts with 68 eligible school districts of which 20 did not meet staffing requirements and 5 refused to participate. Then there were another 15 “Not selected for recruitment”. It is - is it not? - simply true that how and why these 15 and not another 15 were left out is “undocumented”? I interpret what you say now as that schools which met staffing requirements and were willing to participate were sequentially recruited until you had 28, leaving 15 left over. Still it is not clear how the order in which schools were approached first was determined. Or was it that all were approached and the first 28 who had the right staff and signalled their willingness were recruited? Still what you have said leads me to think the doubts cast by the term “undocumented” may not be appropriate and have taken this out.

The issue here is the implicit implication of “left out” versus “recruitment ceased once the targeted sample size was reached.” “Left out” implies we chose not to include them for “undocumented” reasons; had they agreed to participate before our sample size was reached, they would have been included in the study (while other sites would have been “left out for undocumented reasons”).

20. You say, “First, we would like to draw the readers’ attention to prior commentaries addressing similar validity issues, including one on the validity of PROSPER findings (Rulison, Feinberg, Gest, & Osgood, 2016; Spoth, Trudeau, Redmond, & Shin, 2008; Spoth et al., 2017b; Spoth, Trudeau, Redmond, & Shin, 2009). We encourage readers to review these relevant commentaries. In particular, we would like to highlight earlier responses concerning the point in the current Findings commentary about issues with the evidence on one of the programs on the PROSPER menu (the Strengthening Families Program for Parents and Youth 10-14). Notably, studies of the SFP 10-14 program that are cited as raising a question about its efficacy have many differences from the studies conducted by our team, including the fact that none followed participants for more than 3 years post-implementation. For example, some of the studies were conducted with younger students among whom there are very low prevalence rates; our studies indicate that differences between intervention and control groups emerge later, when given types of substance use are more normative. Especially to the point, as noted previously, the summary of the findings from other studies ignores the full range differences from the original trial—in designs, samples, intervention adaptations, and country and cultural contexts—factors that should be considered in a balanced discussion of the generalizability of findings and the current efficacy of the program with rural US populations and elsewhere. We would encourage readers to search out the original articles and commentaries for a more balanced and complete view. Similar to the first key validity point about considering all of the relevant data, in balance, there are additional articles providing results from both of the referenced earlier research projects (e.g., into young adulthood) that were not cited in

the commentary or linked documents."

Have now included your concerns about the adequacy of some of the studies of the Strengthening Families Program.

The Rulison debate I was aware of but did not want to get drawn into the diffusion issue. Only so much can be presented to readers and we are already presenting much, and this is a process issue, not an outcomes one.

Regarding Spoth, R., Trudeau, L., Redmond, C., & Shin, C. (2008) and Spoth, Trudeau, Redmond, & Shin, 2009, we have added references to this debate.

Spoth et al., 2017b
I was not yet aware of this (rather persuasive) rejoinder and have added reference to it.

We are pleased that you took the time to review it, especially since we saw your acknowledgements of inputs from Dr. Gorman in your original review.

21. You say, "Second, we agree that it is appropriate to consider application 'beyond a reasonable doubt' criteria to scientific evaluations. Our concern is the lack of specificity about how these criteria are operationally defined and applied, especially in the case of an article publishing findings from a longitudinal prevention trial originating in 2002 and conducted through the present, over a period of 15 years. In this vein, our view is that there is a need to apply appropriate rigor to the reporting and interpretation of findings – balancing 'beyond a reasonable doubt' with 'the preponderance of evidence' to arrive at the most reasonable conclusion. Although we recognize that opinions on how to operationalize such criteria may differ, the point of the research endeavour should be to advance the science and produce useful knowledge."

    Agree - the preponderance of the evidence is certainly to be taken into account, regardless of p values. One of our first statements in the commentary was: "Relative to control communities, on every single measure published to date, children recruited to the trial in PROSPER communities were less likely to develop the substance use patterns the programme was aiming to prevent. On the balance of probabilities, it is likely that there were preventive impacts."

22. You say, "Finally, it is worth noting that the lead PROSPER investigators recently wrote a commentary that summarized guidelines for constructive criticism (Spoth et al., 2017b). A number of these guidelines are of relevance in the case of the present commentary, including careful attention to methodological detail across multiple reports from large programs of research."

    That article is now referred to in the commentary.

Mike Ashton
15 November 2017

A copy of the original *Drug & Alcohol Findings* review article to which this commentary responds is available upon request.

Additional Citations

Riggs, N.R., Chou, C., Pentz, M.A., 2009. Preventing growth in amphetamine use: long-term effects of the Midwestern Prevention Project (MPP) from early adolescence to early adulthood. *Addiction, 104*, 1691-1699.

Riggs, N.R., Pentz, M.A., 2009. Long-term effects of adolescent marijuana use prevention on adult mental health services utilization: The Midwestern Prevention Project. *Substance Use & Misuse.* Online Journal: https://doi.org/10.1080/10826080902809691

Oesterle, S., Hawkins, J.D., Kuklinski, M.R., Fagan, A.A., Fleming, C., Rhew, I.C., Brown, E.C., Abbott, R.D., Catalano, R.F., 2015. Effects of Communities That Care on males' and females' drug use and delinquency 9 years after baseline in a community-randomized trial. *American Journal of Community Psychology, 56,* 217–228.